Network Sampling for Rare Trait Inference

Iris Shimizu and Monroe Sirken
National Center for Health statistics

Abstract[1]

Network sampling is a technique which may be used to increase the efficiencies of sample surveys aimed at producing estimates about rare populations. Network sampling can be applied in all sample surveys when a multiplicity counting rule is used for linking individual observation units to multiple selection units. The number of selection units linked to an observation unit is the multiplicity of that unit. In a conventional survey, each observation unit is countable at one, and only one, selection unit to which it is linked by a unitary counting rule, a rule for which the "multiplicity" is one. Because multiplicity rules spread individual observation units over more selection units than a unitary rule, the network sampling technique can be more efficient than a conventional sample of rare populations and/or elusive populations which are difficult to survey at their usual residences. This paper discusses the network sampling methodology and some of its advantages and disadvantages.

Keywords: Sampling design, multiplicity estimators

1. Introduction

The Webster dictionary defines "rare" as "seldom occurring or found: Uncommon." The property of "seldom occurring" or "uncommon" frequently means there is no list of the rare population which can be used as a sampling frame in a survey to make inferences about that rare population.

When the frame for a targeted population is not available but the targeted population is linked to a second population for which there is a frame that can be used to select a probability sample, an indirect sampling method can be used to make inferences about the targeted population. Links between the two populations are used to develop weights that can provide for unbiased estimators and variance estimates.

Network sampling is an indirect sampling method. Network sampling is a technique that assures unbiased estimates when the same targeted population units are eligible to be counted at (linked to) multiple selection units in the survey (Sirken, 1998). Technically speaking, network sampling is not really a sampling technique, because it does not specify rules for selecting a sample. It can be applied to all sampling designs when a survey counting rule links multiple selection units to the same targeted population unit.

In the following, the term "population" will be omitted from "targeted population unit" when omission of that term does not risk confusion. Also, to facilitate discussion in the following, the "network" for a targeted unit is defined as the group of selection units which are eligible to report that targeted unit under the survey counting rules. The multiplicity for a targeted unit is the number of selection units which comprise that targeted unit's network. An individual selection unit may be linked to more than one network, but each targeted unit has only one network.

In the following, the history of network sampling is reviewed in Section 2. Network sampling features are described in Section 3. Sampling errors, measurement errors, and survey costs associated with network sampling are discussed in Sections 4, 5, and 6, respectively. Section 7 gives a summary.

2. History

Network sampling emerged as a new technique in the early 1960s in response to estimation problems involving a sample survey of medical providers designed to estimate prevalence of cystic fibrosis – a relatively rare genetic disease of children. In that survey, multiple providers reported the same patients. Without adjustment, conventional estimators would have counted the same patient as many times as they were reported by different providers and the prevalence estimates of cystic fibrosis would have been biased.

To address the situation, Birnbaum and Sirken (1965) derived three unbiased estimators, each of which addressed the effect of multiplicity on selection probabilities of reported patients. The estimators differed with respect to the kinds of information required about network sizes of targeted units counted in the survey. Of these three estimators, the multiplicity estimator was simplest and most robust and is now generally used whenever network sampling is used.

Multiplicity estimators count every report of a targeted population unit, including duplicate reports of the same

[1] The opinions expressed in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics.

individual targeted unit. They weight each report by the inverse of the multiplicity of the reported unit. The estimators do not require matching the reported units for duplications. They may also be unbiased if every member of the population targeted in the survey is linked to at least one selection unit.

Initially, network sampling was used with surveys for which multiple selection units appeared to be unavoidably linked to the same population unit. Many of these were establishment surveys involved with estimating population prevalence based on counts of individuals having transactions with establishments whose constituents overlap.

It was not until the 1970s that network sampling was applied as a deliberate strategy to improve design efficiency. Sirken (1970) showed that network samples could increase survey yields and decrease sampling errors, especially in population surveys of relatively rare events. He proposed network sampling in household surveys by linking individuals to households of relatives and others with whom they had well defined relationships and who could serve as good proxy respondents. Network sampling with kinship relationships was used in several health surveys to measure such things as: diabetes prevalence, cancer prevalence, births and marriages, recent decedents, the Jewish population. Subsequently, it became apparent that network sampling had potential for reducing measurement error. For example, network sampling using kinship counting rules was used in a pretest of a post-enumeration population survey to evaluate coverage in the decennial Census.

More recently, Sirken et al (1995) demonstrated the utility of network sampling in the Linked Establishment/Population (LEP) survey to estimate the volume of transactions between health care providers and patients. These surveys are conducted in two phases. In phase I, a population survey is conducted and the respondents are asked to report the establishments (for example: medical providers, churches, political parties) with whom they have transactions. In phase II, a survey is conducted at a sample of the establishments reported in phase I to collect data about the transactions of the establishments with all households. In LEP surveys, the network counting rule says that an establishment's transactions are countable at every household whose residents have transactions with the establishment. For example in phase I, if a LEP survey asks about diabetics and their transactions with physicians and there are only two diabetics in the population who see a single physician and those two live in separate households, then the transactions of the two diabetics are countable at both households of those diabetics because those two have transactions with the same provider. To date, the theory for LEP surveys has only considered estimation for the number of transactions between household populations and establishments. It would be possible to use a LEP survey to estimate the prevalence of a rare population if the counting rule is such that each member of the targeted population has transactions with one establishment. In addition to measuring transaction volume, the LEP surveys have application in establishment surveys when free-standing lists of all establishments are not available or not adequate for use as sampling frames, a situation that is expected for rare establishments (Sirken and Shimizu, 2005).

3. Network sampling features

In addition to sample selection procedures (simple random, stratified, cluster sampling, etc.), three survey design features are key to network sampling designs. These are the survey counting rules, the estimators, and the responding rules.

3.1 Counting rules

The counting rule is an essential design feature of surveys that specifies the conditions for linking targeted population units to selection units at which they can be enumerated (Sirken 1975). A targeted unit may be linked to more than one selection unit but it has only one network.

Conventional sampling applies when the survey has a unitary counting rule that links each targeted unit to exactly one selection unit at which it is enumerable. The rule that links each person to his/her usual place of residence is a unitary counting rule that is widely used in household surveys.

Network sampling applies when the survey uses a multiplicity counting rule that allows the same targeted unit to be enumerable at multiple selection units. An example of a multiplicity rule is the self/sibling rule in which a person reports himself and his/her siblings.

3.2. Survey Estimator

Estimators are algebraic algorithms for weighting enumerated targeted units to estimate population parameters. The estimator generally used with network sampling is a multiplicity estimator which was described earlier. For an illustration of these estimators, assume a population of N targeted units $I = \{I_1, \ldots, I_\alpha, \ldots I_N\}$ in L selection unit households $H = \{H_1, \ldots, H_i,$

…, $H_L\}$. The network estimate of N based on a simple random sample of ℓ selection units is then

$$\hat{N} = \frac{L}{\ell} \sum_i^{\ell} \lambda_i \ ,$$

where $\lambda_i = \sum_\alpha W_{\alpha i}$ is the weighted sum of the targeted units countable at selection unit H_i and $W_{\alpha i}$ is the network weight assigned to I_α when I_α is counted at H_i. The network estimator is unbiased if, and only if, the sum of weights across all selection units is 1 for each targeted unit, that is if

$$\sum_i W_{\alpha i} = 1 \ , \ \alpha = 1, \ldots, N.$$

The multiplicity estimator assigns the network weights

$$W_{ai} = S_{\alpha i} \big/ S_\alpha \ , \ \alpha \ = 1\ldots, N, \text{i}=1, \ldots, L$$

where $S_{\alpha i}$ is the number of times that I_α is linked to H_i, and S_α is the total number of links I_α has with all selection units.

The conventional estimator is a special case of the network estimator in which S_α =1 and, thus, weight $W_{\alpha i}$ = $S_{\alpha i}$ for all targeted units (α = 1, …, N) and all selection units (i= 1, …, L). That is:

$$S_{\alpha i} = \begin{cases} 1 & \text{if } H_i \text{ is the one H to which } I_\alpha \text{ is linked} \\ 0 & \text{otherwise.} \end{cases}$$

Note that multiplicity estimators require network weights for the I_α enumerated at sample selection units and no others. Hence, it is cost effective to collect the information needed for the multiplicity weights from the selection units at which the targeted units are enumerable.

3.3 Respondent rules

Respondent rules specify the selection units that are eligible to provide information about the targeted population units enumerated in the survey. Three kinds of information are collected.

- Eligibility information identifies the targeted units countable at the selection unit in compliance with network counting rules.
- Topic information is data, if any, about the countable targeted units for the variables of interest in the survey.
- Network information is used to determine the network weights.

For illustration, assume a diabetes prevalence household survey that uses a self/sibling counting rule and a multiplicity estimator. The eligibility information is a list of diabetic residents in the sampled household and diabetic non-resident siblings. The topic information is data for the survey variables of interest about each of the reported diabetics (for example, these may be age, race, sex, etc.). The network information for each diabetic is the total number of siblings in the diabetic's family.

4. Design effects

Research has been done on design effects of the network samples which compared the variances of network sample estimates with those of conventional sampling estimates based on equivalent sample sizes. Sirken (1970) showed that the difference between sampling variances of the multiplicity and conventional estimators depends on configurations of linkages between selection and targeted units that are formed by conventional and multiplicity counting rules. Network sampling is not necessarily better than conventional sampling for all linkage con-figuration but is likely to be better than conventional sampling for some linkage configurations.

Network sampling is better than conventional sampling when multiplicity counting rules produce specified kinds of linkage configurations. One such configuration is when none of the selection units is linked to multiple targeted units by the multiplicity rules.

The results of research about design effects of network sampling inspired the use of network sampling in household sample surveys of rare populations. Network sampling for rare populations is useful when the multiplicity rules satisfy three conditions.

1. The multiplicity of every individual is equal to or greater than one. As indicated earlier, this condition is necessary in order to avoid coverage errors.
2. Individuals are linked to households that are able and willing to report the multiplicities of the individuals and the variables of interest for the individuals. This is necessary to control response errors.
3. The distribution of the multiplicities has a large mean (say, greater than or equal 2) and small variance. Such a distribution in the multiplicities

improves sampling variances. This condition is usually satisfied when the counting rules distribute the individuals as uniformly as possible across households.

5. Measurement errors

Counting rules that reduce sampling errors may adversely affect non-sampling errors and vice versa. However, even when network sampling increases response biases, it can potentially still reduce mean squared errors when sample sizes and prevalence rates are sufficiently small. For example, this was demonstrated by the Survey Research Laboratory of U. of Illinois which conducted a pilot household survey of cost of cancer care in which it investigated the sampling error and response bias of three counting rules in estimating cancer prevalence (Czaja et al. 1986). The sample was seeded with households from the Illinois cancer registries, unknown to the field staff. The three counting rules by characteristics of patients and respondents were:

- Rule 1: A unitary rule which linked cancer patients to their own residences.
- Rule 2: A multiplicity rule which linked cancer patients to their own residences and those of their siblings.
- Rule 3: A multiplicity rule which linked cancer patients to their own residence and those of their children.

Relative to the unitary rule 1, multiplicity rules 2 and 3 reduced sampling errors of unitary rule by about 40% and 15%, respectively. Rule 2 doubled the response bias of rule 1 while rule 3 increased the response bias by more than a third. Even so, for estimating cancer prevalence rate of about 2%, rule 2 was more efficient than Rule 1 for samples of 600 and Rule 3 was more efficient than Rule 1 for samples of 2,100.

6. Survey Costs

Because network sampling increases survey yields, respondent households can report more individuals than in conventional sampling. Network sampling also requires multiplicities for the reported individuals, which is information not required in conventional sampling where the multiplicity is always 1. Interviews are, thus, longer and the cost per household is greater than that in conventional sampling. Hence, to be cost effective, network sampling needs to be more efficient (require smaller samples) than conventional sampling. This is most likely to occur when conventional sampling is subject to large sampling errors and/or non-sampling errors.

As mentioned above, research has shown that network sampling reduces sampling errors when none of the selection units are linked to multiple targeted units under the multiplicity rule. Experience has shown that network sampling errors are usually better when the average number of targeted units per selection unit is small. These two conditions are frequently satisfied by rare populations.

7. Summary

Flexibility with respect to network size provides network sampling with design options that are potentially useful in addressing survey design problems that challenge conventional sampling, especially when multiple selection units are unavoidably linked to the same targeted population units. The technique can also be used as a strategy to improve survey efficiency when conventional sampling results in large sampling and/or measurement errors.

Network sampling is better (reduces sampling errors) than conventional sampling in household surveys of rare populations, because opportunities are enhanced for distributing individuals more uniformly over households by multiplicity rules than by unitary rules.

On the other hand, network sampling comes with a price. The network estimator requires the multi-plicities of targeted units reported in the sample survey, which is information not required by the conventional sampling estimator because, then, the multiplicity is one for every unit. The additional data collection adds to the survey costs, even though the multiplicities are usually reported by the same respondents who reported the targeted units. More serious than added survey costs of collecting the multiplicities is the risk of reporting errors in reporting the multiplicities.

Network sampling is not a perfect solution for dealing with survey design problems that challenge conventional sampling, but when used judiciously and selectively, it has potential for improving survey design efficiency. In particular network sampling can be useful in surveys of rare populations.

References

Birnbaum ZW and Sirken MG (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. National Center for Health Statistics. Vital Health Stat 2(11).

Czaja RF, Snowden CB, and Casady RF (1986). Reporting Bias and Sampling Errors in Surveys of

Rare Populations Using Multiplicity Counting Rules. *JASA* 81: 411-419.

Sirken, MG (1970). Household Surveys with Multiplicity. *JASA* 65: 257-266.

Sirken MG (1975). The Counting Rule Strategy in Sample Surveys. *1975 Proceedings of the Social Statistics Section, American Statistical Association.* 340-342.

Sirken MG (1998). Network Sampling. In Armitage P and Colton T, eds. *Encyclopedia of Biostatistics.* West Sussex, England. John Wiley and Sons Ltd. 2977-2986.

Sirken M, Shimizu I, and Judkins D. (1995). The Population Based Establishment Surveys. *1995 Proceedings of the Survey Research Section, American Statistical Association*. pp 470-473.

Sirken M and Shimizu I (2005). Establishment Surveys with Population Survey-Generated Sampling Frames. In Armitage P and Colton T, eds. *Encyclopedia of Biostatistics, Second Edition*, West Sussex, England. John Wiley and Sons Ltd., 1750-1755.